An Empirical Study on the Fairness of Pre-trained Word Embeddings

Emeralda Sesari, Max Hort and Federica Sarro
Department of Computer Science
University College London
emeralda.sesari.20@alumni.ucl.ac.uk,
{max.hort.19, f.sarro}@ucl.ac.uk

Abstract

Pre-trained word embedding models are easily distributed and applied, as they alleviate users from the effort to train models themselves. With widely distributed models, it is important to ensure that they do not exhibit undesired behaviour, such as biases against population groups. For this purpose, we carry out an empirical study on evaluating the bias of 15 publicly available, pre-trained word embeddings model based on three training algorithms (`GloVe`, `word2vec`, and `fastText`) with regard to four bias metrics (WEAT, SEMBIAS, DIRECT BIAS, and ECT). The choice of word embedding models and bias metrics is motivated by a literature survey over 37 publications which quantified bias on pre-trained word embeddings. Our results indicate that `fastText` is the least biased model (in 8 out of 12 cases) and small vector lengths lead to a higher bias.

1 Introduction

Word embeddings are a powerful tool and are applied in variety of Natural Language Processing tasks, such as text classification (Aydoğan and Karci, 2020; Alwehaibi and Roy, 2018; Jo and Cinarel, 2019; Bailey and Chopra, 2018; Rescigno et al., 2020) and sentiment analysis (Araque et al., 2017; Rezaeinia et al., 2019; Fu et al., 2017; Ren et al., 2016; Tang et al., 2014). However, analogies such as "Man is to computer programmer as woman is to homemaker" (Bolukbasi et al., 2016a) contain worrisome biases that are present in society and hence embedded in language. In recent years, numerous studies have attempted to examine the fairness of word embeddings by proposing different bias metrics (Caliskan et al., 2016; Garg et al., 2018; Sweeney and Najafian, 2019; Manzini et al., 2019; Dev et al., 2019), and comparing them (Badilla et al., 2020).

The quality of word embedding models differs depending on the task and training corpus used. Due to the relatively expensive costs, constructing large-scale labelled datasets is a huge barrier for NLP applications, notably for syntax and semantically related tasks (Qiu et al., 2020). Recent research has shown that by using pre-trained word embedding models, trained on a large corpus, considerable performance gains on various NLP tasks can be achieved (Qiu et al., 2020; Erhan et al., 2010). A number of studies (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2017) have published these embeddings learned from large text corpora which are versatile enough to be used in a variety of NLP tasks (Li and Yang, 2018). Despite their widespread use, many researchers use word embeddings without performing an in-depth study on their characteristics; instead, they utilised default settings that come with ready-made word embedding toolkits (Patel and Bhattacharyya, 2017). On top of that, these pre-trained models are susceptible to inheriting stereotyped social biases (e.g., ethnicity, gender and religion) from the text corpus they are trained on (Caliskan, 2017; Garg et al., 2018; Vidgen et al., 2021) and the researchers building these models (Field et al., 2021).

Moreover, word embedding models are sensitive to a number of parameters, including corpus size, seeds for random number generation, vector dimensions, etc. (Borah et al., 2021). According to Levy et al. (2015) changes in parameters, are responsible for the majority of empirical differences between embedding models. As a result, there has been an increasing interest among researchers to investigate the impact of parameters on word embedding model properties (e.g., consistency, stability, variety, and reliability) (Borah et al., 2021; Chugh et al., 2018; Dridi et al., 2018; Hellrich and Hahn, 2016; Pierrejean and Tanguy, 2018; Wendlandt et al., 2018; Antoniak and Mimno, 2018). However, much uncertainty still exists about the relation between word embedding parameters and its fairness. With the in-depth investigation of fair-

Proceedings of the The 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 129 - 144

ness, we hope that this research will lead to a more directed and fairness-aware usage of pre-trained word embeddings. Therefore, this study investigates the performance of pre-trained word embedding models with respect to multiple bias metrics. Furthermore, the impact of each pre-trained word embedding model's vector length on the model's fairness is explored. We investigate 15 different scenarios in total as a combination of model, training corpus, and parameter settings. We make the scripts used to determine the fairness of pre-trained word embedding models publicly available.[1]

Bias statement. Word embeddings are used to group words with similar meanings (i.e., generalise notions from language) (Goldberg and Hirst, 2017). However, word embedding models are prone to inherit social biases from the corpus they are trained upon. The fundamental concern is that training a system on unbalanced data may lead to people using these systems to develop inaccurate, intrinsic word associations, thus propagating biases (Costa-jussà and de Jorge, 2020). For example, stereotypes such as $man : woman :: computer\ programmer : homemaker$ in `word2vec` trained on news text can be found (Bolukbasi et al., 2016a). If such an embedding is used in an algorithm as part of its search for prospective programmers, documents with women's names may be wrongly downweighted (Jurafsky and Martin, 2020).

Our research helps practitioners to make an informed choice of fair word embedding models, in particular pre-trained models, for their application with regards to intrinsic biases (i.e., gender, race, age).

2 Background

It has been discovered that word embeddings do not only reflect but also have the tendency to amplify the biases present in the data they are trained on (Wang and Russakovsky, 2021) which can lead to the spread of unfavourable stereotypes (Zhao et al., 2017). The implicit associations which are a feature of human reasoning are also encoded by embeddings (Greenwald et al., 1998; Caliskan et al., 2016). Using the Implicit Association Test (IAT), Greenwald et al. (1998) reported that people in the United States demonstrated to link African American names with bad connotations more than European American names, female names with art related words and male names with math related words. In 2016, Caliskan et al. (2016) used `GloVe` vectors and cosine similarity to recreate IAT and discovered that African American names like *Jamal* and *Tamika* showed higher cosine similarity with unpleasant words like *abuse* and *terrible*. On the contrary, European American names such as *Matthew* and *Ann* had a greater cosine similarity with pleasant terms such as *love* and *peace*. These are an example of *representational harm* where a system causes harm that is demeaning some social groups (Blodgett et al., 2020; Crawford, 2017).

In the context of word embeddings, it is not only of importance to show that bias exists, but also to determine the degree of bias. For this purpose, bias metrics can be used. Bias metrics can be applied either to a single word, a pair of words, or an entire list of words. Percent Male Neighbours (PMN) (Gonen and Goldberg, 2019) is a bias metric that operates on a single word, where one could see the percentage of how many male-gendered words surrounded a target word. For instance, Badilla et al. (2020) discovered that using PMN, 16% of the words around *nurse* are male-gendered words. However, when *engineer* is the target term, 78% of words surrounding it are male-gendered.

Moreover, Bolukbasi et al. (2016a) sought to measure bias by comparing the embeddings of a pair of gender-specific terms to a word embedding. The authors introduced DIRECT BIAS, in which a connection is calculated between a gender neutral word (e.g., *nurse*) and an obvious gender pair (e.g., *brother* − *sister*). They also took into account gender-neutral word connections that are clearly derived from gender (i.e., INDIRECT BIAS). For instance, female associations with both *receptionist* and *softball* may explain why the word *receptionist* is significantly closer to *softball* than *football*.

Similarly, SEMBIAS (Zhao et al., 2018) also uses word pairs to evaluate the degree of gender bias in a word embedding. SEMBIAS identifies the correct analogy of *he*−*she* in a word embedding according to four pairs of words: a gender definition word pair (e.g., *waiter* − *waitress*), a gender-stereotype word pair (e.g., *doctor* − *nurse*) and two other pairs of words that have similar meanings (e.g., *dog* − *cat*, *cup* − *lid*).

In addition, Word Embedding Association Test (WEAT) (Caliskan et al., 2016; Sweeney and Najafian, 2019) determines the degree of association

[1] `https://figshare.com/s/23f5b7164e521cf65fb5`

between lists of words (target and attribute words), to automatically assess biases emerging from word embeddings. A target word set is a collection of words that represent a specific social group and are used to assess fairness (e.g., *Muslims*, *African American*, *men*). While an attribute word set is a set of words denoting traits, characteristics, and other things that can be used to show a bias toward one of the targets (e.g., *career* vs *family*).

Another significant aspect of these metrics is that there is lack of a clear relationship between them (Badilla et al., 2020). They function with diverse inputs, resulting in incompatibility between the outputs. As a result, a number of studies began to examine the use of word embedding fairness frameworks, such as Embeddings Fairness Evaluation Framework (WEFE) (Badilla et al., 2020) and Fair Embedding Engine (FEE) (Kumar and Bhotia, 2020).

3 Paper Selection

The aim of paper selection is to gather published work that refers to word embedding models and metrics used to evaluate the fairness of word embeddings. Following that, we choose the most commonly used pre-trained word embedding models and bias metrics to support our experiments. Due to the scope and recent emergence of this topic, we conduct a comprehensive literature review according to guidelines by Kitchenham (2004). The selection starts with searching for the relevant publications and then extracts pertinent information. Below, we discuss our search methodology in detail, starting with preliminary search, defining keywords, repository search, followed by selecting relevant papers based on the inclusion criteria and snowballing.

3.1 Search Methodology

3.1.1 Preliminary Search

A preliminary search was carried out prior to systematically searching online repositories. This search is particularly useful in understanding the field and the extent to which fairness of word embeddings is covered in previous studies. The results were used to determine keywords (Table 1) which then guided the repository search.

3.1.2 Repository Search

Following the preliminary search, a search on the online libraries of six widely known repositories, namely, ACM Digital Library, arXiv, IEEE Xplore, Google Scholar, ScienceDirect, and Scopus, was conducted. Notable, Google Scholar contains publications from the ACL Anthology. [2] The search took place on 8 June, 2021. Unlike Hort et al. (2021), this search was not restricted by year. However, prior to commencing the search, an agreement was reached on the specific data field used in the search of each repository, thereby limiting it to the specific parts of a document record. Appendix A shows the data fields used during this search. In particular, the repository search investigates the combination of each keyword pair among the two categories (as shown in Table 1).

Category	Keywords
Word embedding model	word embedding, word embedding model, pre trained word embedding model, pre-trained word embedding
Bias or Fairness	fairness, fairness metrics, bias, bias metric

Table 1: Keywords defined from the preliminary search.

3.1.3 Selection

We evaluate the following inclusion criteria to ensure that the publications found during the search are relevant to the topic of fairness of pre-trained word embeddings:

- The publication investigates the fairness of pre-trained word embeddings;
- The publication describes the specific metric or measurement of assessing the fairness of word embeddings;
- The studied metrics are intrinsic, i.e., measuring bias directly in word embedding spaces (Goldfarb-Tarrant et al., 2021a);
- The studied word embeddings are in English.

To determine if the publications met the inclusion criteria, we manually analysed each publication following the process of Martin et al. (Martin et al., 2017):

1. **Title:** To begin, all publications with titles that clearly do not meet our inclusion criteria are omitted;
2. **Abstract:** Second, every title-selected publication's abstract is examined. At this stage, publications whose abstracts do not fit the inclusion requirements are eliminated;

[2] https://aclanthology.org/

	ACM	arXiv	GS	IEEE	SD	Scopus
Hits	21	94	19	64	30	58
Title	18	88	19	24	8	47
Abstract	12	84	19	12	2	34
Body	2	28	3	0	0	4
Total	37					

Table 2: Repository search results.

3. **Body:** Publications that have passed the first two steps are then reviewed in full. In case the material does not meet the inclusion criterion or contribute to the survey, they are excluded.

The number of publications gathered from online repositories was reduced by removing the duplicates and applying both the aforesaid process and inclusion criteria. The first and second author participated in this process, and differences were discussed until an agreement was made. In the section 3.3, we investigate the set of relevant publications as the result of this paper selection.

3.1.4 Snowballing

After selecting a set of relevant papers from the repository search, one level of backwards snowballing (Wohlin, 2014) was done to examine their references. It entails reviewing the bibliographies of selected publications, determining whether they are relevant, and adding them to the list.

3.2 Selected Publications

The results of the repository search are shown in Table 2. The first column contains the six online repositories mentioned in Section 3.1.2, in which Google Scholar is abbreviated with GS and Science Direct is abbreviated with SD. The overall number of publications found using the keywords (Table 1) and filters (Appendix A) provided is shown in the first row, while the number of relevant publications filtered based on the paper title, abstract, and body is shown in the last three rows. In addition to the 37 publications retrieved from the repository search, we considered 7 publications from a preliminary search and 1 additional from snowballing.

3.3 Results

Through a comprehensive search, this study looked at the current literature on the fairness of pre-trained word embeddings. In total, we compiled a list of 23 distinct bias metrics that were used to evaluate the fairness of pre-trained word embeddings. It is worth noting that a publication might use multiple pre-trained models and bias metrics (Schlender and Spanakis, 2020; Spliethöver and Wachsmuth, 2020; Friedrich et al., 2021; Wang et al., 2020; Vargas and Cotterell, 2020; May et al., 2019; Dev et al., 2020). The more detailed explanation of the result is discussed in the following sections.

3.3.1 The most frequently used pre-trained static word embedding model

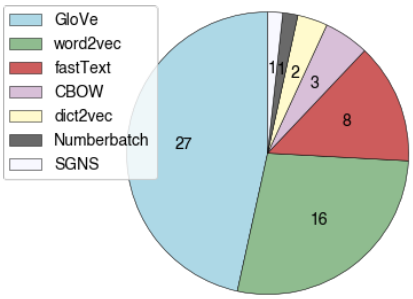
One of the goals of the paper selection was to extract the most relevant pre-trained word embedding models from the many that have been studied. While recent research on contextual embeddings has proven immensely beneficial, static embeddings remain crucial in many situations (Gupta and Jaggi, 2021). Many NLP applications fundamentally depend on static word embeddings for metrics that are designed non-contextual (Shoemark et al., 2019), such as examining word vector spaces (Vulic et al., 2020) and bias study (Gonen and Goldberg, 2019; Kaneko and Bollegala, 2019; Manzini et al., 2019). Furthermore, according to Strubell et al. (2019), the computational cost of employing static word embeddings is often tens of millions of times lower than the cost of using contextual embedding models (Clark et al., 2020), which is significant in terms of NLP models financial and environmental costs (Strubell et al., 2019). Therefore, we focus our proceeding investigation to static models. The number of papers that have looked into fairness on a pre-trained static word embedding model is shown in Figure 1a.

It is apparent from this chart that pre-trained model `GloVe` is the most popular in this research field. The second and third most frequently used models are `word2vec` and `fastText`, respectively. Appendix C Table 7 lists all seven distinct pre-trained word embedding models we found during our search.

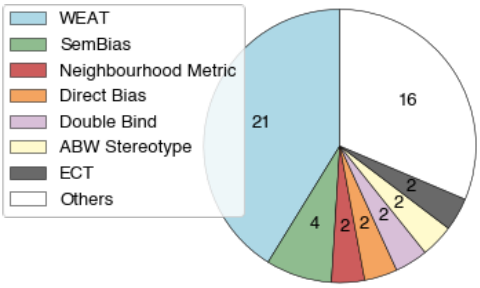
3.3.2 The most frequently used bias metrics

The paper selection's next aim was to select the most commonly used bias metrics from among the numerous that have been used to examine the fairness of a pre-trained word embedding model. 23 metrics were gathered and sorted based on the number of papers that used them.

To minimise space, bias metrics that have only been utilised in one study have been labelled as *Others*. As can be seen from Figure 1b, WEAT is by far the most prevalent bias metric, with 21 out of 32 of the publications using it to quantify bias



(a) Collected pre-trained static word embedding models.



(b) Collected bias metrics.

Figure 1: Publications investigating fairness on pre-trained static word embedding model

in pre-trained word embeddings. The second most used metric is SEMBIAS which was used by 4 out of 32 publications. In addition, we found 5 bias metrics which were used by 2 out of 32 publications: NEIGHBOURHOOD METRIC, DIRECT BIAS, DOUBLE BIND, ABW STEREOTYPE and ECT. Appendix C Table 8 lists the detailed information for these metrics including sixteen other metrics that were only utilised in one research.

4 Empirical Study Design

4.1 Research Questions

The answer to the following research questions is sought to raise awareness on biased behaviour in commonly used pre-trained word embedding models:

RQ1 How do pre-trained word embeddings perform with respect to multiple fairness measures?

A series of experiments were carried out to better understand how pre-trained word embeddings perform when subjected to different fairness measures. The most commonly used bias metrics (WEAT, SEMBIAS, DIRECT BIAS, and ECT) were used to assess the fairness of the three most popular pre-trained embeddings: `GloVe`, `word2vec`, and `fastText` (see Sections 3.3.1 and 3.3.2). Fairness here refers to the absence of bias in a word embedding model; if the bias is high, the degree of fairness is low, and vice versa. Hence, we examined the most fair embedding after the bias values were acquired.

RQ2 How does the vector affect word embedding fairness?

To investigate the effect of vector length on the fairness of pre-trained word embedding models, we compare embeddings trained on the same corpus. Therefore, we investigate `GloVe` Twitter and `GloVe` Wiki Gigaword to determine the effect.

4.2 Design Choice

4.2.1 Pre-Trained Embeddings

We performed experiments using publicly available pre-trained word embeddings. Please refer to Table 3 for the details about the embeddings. These embeddings are provided by the three most used embedding models described in Section 3.3.1.

GloVe was trained under three different corpora, resulting in 10 pre-trained word embeddings: four embeddings from 2 billion tweets of Twitter corpus, four embeddings from 6 billion tokens of Wikipedia and Gigaword corpus, two embeddings each from 42 billion and 840 billion tokens of Common Crawl corpus. Pre-trained embeddings trained on Twitter and Wikipedia + Gigaword corpus have varying dimensionalities (i.e., vector length). We also investigated a pre-trained **word2vec** embedding model, which was trained on 3 billion tokens on a Google News corpus with a vector length of 300. Finally, we evaluated four pre-trained embeddings from **fastText**, each with and without subword information, on 16 billion tokens from Wikipedia + UMBCWeb Base + statmt.org News and 600 billion tokens from Common Crawl.

4.2.2 Bias Metrics

We evaluated the fairness of pre-trained word embeddings stated in Section 4.2.1 by focusing on 4 most frequently used and publicly available bias metrics: WEAT, SEMBIAS, DIRECT BIAS, and ECT. To ensure that we measure bias correctly, we focus our evaluation on the metrics that have been used at least twice and are implemented by existing fairness frameworks (e.g., WEFE, FEE). We explain each of these measures below.

Model	Corpus	Token	Vocabulary	Format	Vector Length	File Size
GloVe	Twitter (2B tweets)	27B	1.2M	uncased	25, 50, 100, 200	1.42 GB
	Wikipedia 2014 + Gigaword 5	6B	400K	uncased	50, 100, 200, 300	822 MB
	Common Crawl	42B	1.9M	uncased	300	5.03 GB
		840B	2.2M	cased	300	5.65 GB
word2vec	Google News	3B	∼100B	uncased	300	1.66 GB
fastText	Wikipedia 2017, UMBC Web Base and statmt.org News	16B	1M	cased	300	2.26 GB
			1M + subword	cased	300	2.26 GB
	Common Crawl	600B	2M	cased	300	4.51 GB
			2M + subword	cased	300	4.52 GB

Table 3: Pre-trained word embeddings learned on different sources provided by `GloVe`, `word2vec`, and `fastText`.

In order to unveil bias, WEAT detects whether there is a difference in the strength of association between the two target sets (X, Y) towards attribute sets (A, B):

$$s(X, Y, A, B) = \sum_{x \epsilon X} s_w(x, A, B) - \sum_{y \epsilon Y} s_w(x, A, B)$$

$$s_w(w, A, B) = mean_{a \epsilon A} cos(\overrightarrow{w}, \overrightarrow{a}) - mean_{b \epsilon B} cos(\overrightarrow{w}, \overrightarrow{b})$$

A and B are attribute sets of identical size. $s(X, Y, A, B)$ computes the test statistic and $s_w(w, A, B)$ calculates the difference in similarity of attribute sets to a word w. We focused only on the degree of bias (i.e., we do not consider the direction of bias) and thus only used absolute bias scores for metrics such as WEAT. We utilised WEFE for WEAT experiments and we applied 7 out of 10 WEAT tests provided by Caliskan et al. (2016). We only selected tests that are concerned with protective attributes concerning human biases (i.e., race, gender, and age). We categorised 7 WEAT tests as: racial bias (T3, T4, and T5); gender bias (T6, T7, and T8); and age bias (T10). Please refer to Appendix B for more information about target and attribute sets.

We also evaluated the degree of bias in pre-trained word embeddings by using the SEMBIAS metric provided in FEE. Zhao et al. (2018) developed this analogous dataset with 20 gender-stereotype word pairs and 22 gender-definitional word pairs, resulting in 440 instances using their Cartesian product. Each instance consists of four-word pairs: a gender definition word pair or Definition (e.g., *waiter − waitress*), a gender-stereotype word pair or Stereotype (e.g., *doctor − nurse*), and two none-type word pairs or None (e.g., *dog − cat*, *cup − lid*). The bias according to SEMBIAS is then measured by iterating over each instance and determining the distance vector of each of the four word pairs. The percentage of times that each word pair type achieves the highest similarity to *he − she* based on their distance vector is measured, with a "Definition" percentage close to 1 is desirable.

We applied DIRECT BIAS (Bolukbasi et al., 2016a) to measure bias with regards to a list gender neutral words N and the gender directions g:

$$DirectBias = \frac{1}{|N|} \sum_{w \epsilon N} |cos(\overrightarrow{w}, g)|^c$$

The parameter c determines how strict the bias measurement is. We conducted the experiment by using DIRECT BIAS that has been implemented in FEE with a 320 profession word list[3] provided by Bolukbasi et al. (2016a) and $c = 1$. Lower DIRECT BIAS scores indicate that a word embeddings is less biased.

The EMBEDDING COHERENCE TEST (ECT) (Dev and Phillips, 2019) computes gender bias based on the rank of the nearest neighbors of gendered word pairs ε (e.g., "she" - "he"). These gendered word pairs, consisting of female and male terms, are averaged, such that two mean embedding vectors m and s remain (one for female terms and one for male terms). Given a list of words affected with indirect bias P, in this case a list of professions proposed by Bolukbasi et al. (Bolukbasi et al., 2016a), the similarity of each word to m and s is determined. The cosine similarities are then replaced by rank order, and given m and s, we receive two rank orders for the words in P. Next, the Spearman Coefficient is calculated once the ranks are compared. For each word pair, ECT is optimised with a Spearman

[3] https://github.com/tolga-b/debiaswe

Coefficient towards 1. Here, we experimented with ECT that has been implemented in WEFE using male and female names as target sets, and professions as attribute set. All word list are available in the ECT online repository.[4]

The measures used in this paper only examine for particular bias types, not all of them. As a result, these measures can only be used to indicate the presence of these specific types of bias and cannot be used to establish the absence of all biases.

5 Empirical Study Results

5.1 RQ1: Fair Pre-trained Word Embeddings

Table 4 reports the bias score obtained from the experiment described in Section 4.1 together with pre-trained embeddings and bias metrics chosen in Section 4.2. Bold bias score indicates the best score of the corresponding measure while arrows next to the measure represent the interpretation of the score: downward arrow means the lower the value, the less biased an embedding is; upward arrow means the higher the score, the less biased an embedding is.

5.1.1 WEAT

The purpose of this experiment is to measure the degree of association between target and attribute words defined by Caliskan (2017) to assess biases emerging from the pre-trained word embeddings. From Table 4, it can be seen that pre-trained `fastText` models resulted in the lowest bias for tests concerned with racial bias, age bias, and gender bias with gendered names involved. `fastText` Wiki News scored the lowest on Test 3 and Test 4, whereas `fastText` Wiki News with subword information scored the lowest on Test 5. `fastText` Wiki News is also the least biased embedding in terms of age bias (Test 10). Interestingly, among all tests with respect to gender bias: Test 6, Test 7, and Test 8, `fastText` only outperforms other models on Test 6, particularly `fastText` that has been trained under Common Crawl corpus with subword information.

Turning now to WEAT tests with respect to gender bias which use male and female terms as the attribute words: Test 7 and Test 8. Closer inspection of the Table 4 reveals that pre-trained embeddings trained with `GloVe` model using Twitter corpus with vector lengths of 200 and 100, outperform other embeddings across the two tests, respectively. Taken together, these results acquired from WEAT tests suggest that `fastText` is the least biased model for 5 out of the 7 WEAT tests.

5.1.2 SEMBIAS

This experiment is aimed at identifying the correct analogy of *he* − *she* in various pre-trained word embeddings according to four pairs of words defined by Zhao et al. (2018). The results obtained from the SEMBIAS experiment can be compared in Table 4. It is expected to have a high accuracy for Definitions and low accuracy for Stereotypes and Nones.

This table is quite revealing in several ways. First, all embeddings trained using `fastText` outperform the other pre-trained embeddings. `fastText` embeddings achieve high semantic, definition scores above 86.8% while keeping stereotypical and none loss to a minimum, below 1% and 3% respectively. Second, among the four embeddings trained with `fastText`, the one trained with Common Crawl is shown to be the least biased. The percentage of Definition, Stereotype, and None predictions achieved by this embeddings are 92.5%, 5% and 2.5%, respectively. Despite the fact that `fastText` Wiki News with subword information embeddings achieved the lowest percentage of None, the Stereotype prediction must not be forgotten. Compared to the Stereotype prediction of `fastText` Common Crawl, `fastText` Wiki News with subword information embeddings correctly classified 0.4% more words as a gender-stereotype word pair, which makes it slightly more biased.

Together, these results provide important insights into how most word pairs in `fastText` pre-trained embeddings are correctly classified as a gender-definition word pair but only few word pairs are correctly categorised as a gender-stereotype word pair and gender unrelated word pairs. Also according to these data, we can infer that `fastText` model trained on the Common Crawl corpus generates the least biased pre-trained word embeddings.

5.1.3 DIRECT BIAS

DIRECT BIAS calculates the connection between gender neutral words and gender direction learned from word embeddings. One unanticipated finding is that the word embeddings generated from the `GloVe` model trained on Wiki Gigaword corpus with vector length 300, is found to be the

[4] https://github.com/sunipa/Attenuating-Bias-in-Word-Vec

Pre-trained Embeddings	WEAT							SemBias			DB↓	ECT↓
	T3↓	T4↓	T5↓	T6↓	T7↓	T8↓	T10↓	D↑	S↓	N↓		
GloVe												
twitter-25	3.753	1.838	1.540	0.818	0.043	0.091	0.329	0.178	0.431	0.391	0.482	0.965
twitter-50	2.564	1.432	1.184	0.736	0.212	0.180	0.354	0.322	0.397	0.281	0.354	0.945
twitter-100	2.189	1.215	1.381	0.654	0.060	**0.004**	0.360	0.508	0.300	0.192	0.140	0.900
twitter-200	1.674	0.918	1.161	0.537	**0.035**	0.063	0.224	0.589	0.278	0.133	0.037	0.916
wiki-gigaword-50	1.893	0.872	1.331	2.317	0.468	0.403	0.320	0.698	0.216	0.133	0.127	0.763
wiki-gigaword-100	1.553	0.971	1.434	1.732	0.366	0.253	0.335	0.750	0.182	0.086	0.135	0.809
wiki-gigaword-200	1.443	0.828	1.114	1.494	0.275	0.335	0.200	0.779	0.168	0.052	0.028	0.769
wiki-gigaword-300	1.279	0.848	1.069	1.319	0.243	0.319	0.212	0.786	0.150	0.064	**0.004**	0.743
common-crawl-42B	1.828	0.894	0.949	0.738	0.260	0.235	0.213	0.805	0.125	0.070	0.627	0.889
common-crawl-840B	1.863	0.971	1.112	1.267	0.199	0.314	0.354	0.830	0.120	0.050	0.450	0.861
word2vec												
google-news-300	0.454	0.453	0.338	1.252	0.225	0.293	0.049	0.827	0.134	0.038	0.082	0.733
fastText												
crawl-300d-2M	0.639	0.328	0.545	0.505	0.221	0.301	0.326	**0.925**	**0.050**	0.025	0.108	**0.692**
crawl-300d-2M-sub	0.902	0.387	0.552	**0.432**	0.268	0.169	0.214	0.868	0.102	0.030	0.083	0.749
wiki-news-300d-1M	0.556	0.266	**0.224**	0.468	0.203	0.163	0.056	0.920	0.055	0.025	0.057	0.752
wiki-news-300d-1M-sub	**0.428**	**0.142**	0.304	0.438	0.198	0.110	**0.026**	**0.925**	0.054	**0.020**	0.035	0.744

Table 4: Bias scores obtained after applying four metrics to several pre-trained word embeddings.

least biased pre-trained embeddings with a score of 0.004. This score confirms that the embeddings have the least gender direction when the gender neutral words being applied to it. Across all bias metrics, DIRECT BIAS is the first one that generates the best score for `GloVe` pre-trained embeddings.

5.1.4 ECT

Similar to WEAT, ECT measures the degree of association between one attribute set and two target sets described in Section 4.2.2. In accordance with WEAT results, a pre-trained `fastText` model was found to be the least biased. Particularly, the `fastText` model that has been trained on the Common Crawl corpus without subword information, has the lowest bias score of 0.692. This score reflects the lack of correlation of the mean vectors distances between the male and female name sets and the occupation words, which result in the smallest presence of bias among all of the embeddings. This result supports evidence from previous experiment with SEMBIAS. The consistency may be due to how both metrics aim to identify a gender bias by utilising occupations as gender neutral words.

5.1.5 Overall

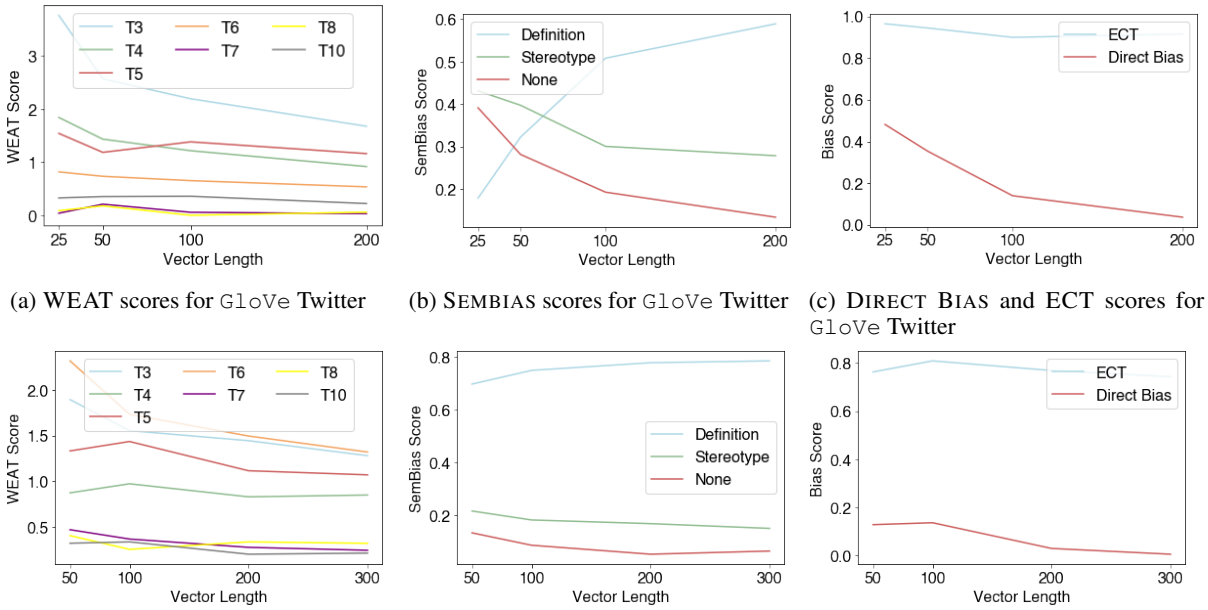
We can infer from these data that `fastText` pre-trained word embeddings perform the best with respect to three of the four most used bias metrics. According to SEMBIAS and ECT scores, `FastText` Common Crawl is the least biased. Using the same corpus but with addition of subword information, the embeddings has the least biased according to WEAT Test 6. Furthermore, `FastText` Wiki News is least biased on WEAT Test 5. In addition, the embeddings has the least bias on WEAT Test 3, Test 4, and Test 10 while including subword information.

5.2 RQ2: Effect of Vector Length on Fairness

The second RQ investigates the impact of parameters on the fairness of pre-trained word embedding models. We conduct experiments to bias in regards to vector length.

Figure 2a and Figure 2d present the results obtained from the analysis of WEAT scores with respect to the vector length. On four of the seven WEAT tests: Test 3, Test 4, Test 6, and Test 7 (after 50 dimension) there is a clear trend of decreasing bias in `GloVe` Twitter with the rise value of vector length (Figure 2a). On the other hand, Figure 2d indicates that the bias in `GloVe` Wiki drops as the vector length increases in four WEAT tests: Test 3, Test 5 (after 100 dimension), Test 6, and Test 7. In summary, 8 from 14 WEAT's findings imply that the greater the `GloVe` Twitter and `GloVe` Wiki dimension, the less biased they are.

Turning now to the analysis on SEMBIAS scores, it is apparent from Figure 2b and Figure 2e that the fairness improves with the increase in the number of dimensions. Note that in SEMBIAS, a high accuracy for Definitions and low accuracy for Stereotypes and Nones are expected. That is why as the dimension rises, the Definition's accuracy increases, but the Stereotype and None's accuracy decreases. Overall, this finding indicates that according to SEMBIAS, words in `GloVe` Twitter and



(a) WEAT scores for GloVe Twitter

(b) SEMBIAS scores for GloVe Twitter

(c) DIRECT BIAS and ECT scores for GloVe Twitter

(d) WEAT scores for GloVe Wiki Gigaword

(e) SEMBIAS scores for GloVe Wiki Gigaword

(f) DIRECT BIAS and ECT scores for GloVe Wiki Gigaword

Figure 2: Bias scores with respect to the vector length.

GloVe Wiki embeddings are more likely to be correctly identified as gender-definition word pair but less likely to be correctly classified as a gender-stereotype word pair and gender unrelated word pairs if they were trained with large vector lengths.

The next analysis of this experimental result is concerned with how the DIRECT BIAS scores would be affected by the vector length. Figure 2c shows that following the increase of vector length in GloVe Twitter, we observe a decrease in the bias score. In Figure 2f, bias score of GloVe Wiki Gigaword increases from lower dimensions 50 to 100 but decreases beyond dimension 100. These results show that from four vector lengths used in each of the two corpora, most of them support the hypothesis that the larger dimension used resulted in smaller presence of gender bias. The rise of bias score of GloVe trained in Wiki Gigaword corpus from 50 to 100 dimension is the only instance that counters our hypothesis.

Lastly, Figure 2c shows a decrease in ECT score as vector length increases in GloVe Twitter only within dimensions of 25, 50, and 100. However, between 100 and 200, the bias score increases by 0.016. In addition, Figure 2f illustrates that the discovery of GloVe Wiki Gigaword in ECT is similar to that in DIRECT BIAS, that the bias increases from lower dimensions 50 to 100 but rapidly declines beyond dimension 100. Six of the eight pre-trained embeddings examined in this investigation support the finding that fairness improves as the number of dimensions increases.

Finally, most observations from the WEAT, SEMBIAS, DIRECT BIAS, and ECT scores indicate evidence for improved fairness in pre-trained word embeddings when the number of dimensions is increased. This result implies that lower dimensionality word embeddings are not expressive enough to capture all word associations and analogies, and that when the bias metric is applied to them, they become more biased than embeddings with larger dimensions.

6 Related Work

There has been a growing interest among researchers to tackle bias in word embeddings, herein we focus on previous work comparing different models and their characteristics.

Lauscher and Glavaš (2019) evaluated embedding space biases caused by four different models and found that GloVe embeddings are biased according to all 10 WEAT tests, while fastText exhibits significant biases only for a subset of tests. This finding broadly supports our finding where all smallest WEAT scores belong to GloVe pre-trained embeddings. However, their focus is different from our as their approach aims at understanding the consistency of the bias effects across

languages, corpora, and embedding models.

Borah et al. (2021) compared the stability of the fairness results to those of the word embedding models used: fastText, GloVe, and word2vec, all of which were trained on Wikipedia. Among the three models, they discovered that fastText is the best stable word embedding model which results in the highest stability for its WEAT results. Badilla et al. (2020) implemented their proposed fairness framework, WEFE, by conducting case study where six publicly available pre-trained word embedding models are compared with respect to four bias metrics (e.g., WEAT, WEAT-ES, RND, RNSB). Consistent with our finding, they discovered that fastText rank first in WEAT.

Lauscher et al. (2019) proposed a general debiasing framework Debiasing Embeddings Implicitly and Explicitly (DEBIE). They used two bias metrics: WEAT Test 8 and ECT to compare the bias of CBOW, GloVe, and fastText trained in Wikipedia. They observed that fastText is more biased than GloVe in both metrics. While this contradicts our observations, their study did not utilise pre-trained models but manually trained them on the same corpus.

Popović et al. (2020) demonstrated the viability of their modified WEAT metric on three classes of biases (religion, gender and race) in three different publicly available word embeddings with vector length of 300: fastText, GloVe and word2vec. Their findings yielded that before debiasing, fastText has the least religion and race bias, while word2vec has the least gender bias. However, one of the study's discoveries opposes our findings where word2vec does not have the least gender bias. This difference may occur given the fact that the authors collected word sets from a number of different literature.

Furthermore, previous work considers the impact of word embedding vector length on the performance and the relation to fairness. Borah et al. (2021) looked at how the length of the vectors used in training fastText, GloVe, and word2vec affected their stability. The models' stability improves as the vector dimensions grow larger. On the other hand, Goldberg and Hirst (2017) found that word embeddings with smaller vectors are better at grouping similar words. This generalisation means that word embeddings with shorter vector lengths have a higher tendency to be biased. The results of our empirical study, obtained using more data and metrics, corroborate the above findings.

Much of the previous research has focused on proposing and evaluating debiasing techniques, modified metrics and fairness frameworks. Therefore, our study makes a major contribution to the research on fairness of word embeddings by empirically comparing the degree of bias of the most popular and easily accessible pre-trained word embeddings according to a variety of popular bias metrics, as well as the impact of vector length involved in the training process to its fairness.

7 Conclusion

The purpose of this study was to empirically assess the degree of fairness exhibited by different publicly available pre-trained word embeddings based on different bias metrics. To this end, we first analysed what are the most used pre-trained word embeddings and bias metrics by conducting a comprehensive literature survey. The results pointed out that the majority of the papers used three word embedding models (namely GloVe, word2vec, and fastText) and four bias metrics (namely WEAT, SEMBIAS, DIRECT BIAS, and ECT). Our results revealed that the most fair of the three pre-trained word embedding models evaluated is fastText. We also found that while using pre-trained embeddings, the influence of vector length on fairness must be carefully considered.

The scope of this study was limited in terms of selecting word list used to apply bias metrics to the word embeddings. We closely examined the earlier studies that may have influenced bias scores. In the future, we need a deeper analysis and explanation of the numerous fairness tendencies discovered in this study, such as the correlation with explicit gender gaps and survey data (Friedman et al., 2019a,b), and the extent to which the embeddings reproduce bias (Blodgett et al., 2021). Moreover, the study could be replicated by not only using pre-trained word embeddings models, but manually training models with different parameters on an identical text corpus. Further study could also be conducted to explore the fairness of contextual word embeddings (e.g., ELMo, Bert), the application bias in word embeddings (Goldfarb-Tarrant et al., 2021b), and bias in word embedding in languages with grammatical gender (Zhou et al., 2019).

Acknowledgments

M. Hort and F. Sarro are supported by the ERC grant 741278 (EPIC).

References

Ali Alwehaibi and Kaushik Roy. 2018. Comparison of pre-trained word vectors for arabic text classification using deep learning approach. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1471–1474.

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

Oscar Araque, Ignacio Corcuera-Platas, J. Fernando Sánchez-Rada, and Carlos A. Iglesias. 2017. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, 77:236–246.

Murat Aydoğan and Ali Karci. 2020. Improving the accuracy using pre-trained word embeddings on deep neural networks for turkish text classification. *Physica A: Statistical Mechanics and its Applications*, 541:123288.

Marzieh Babaeianjelodar, Stephen Lorenz, Josh Gordon, Jeanna Matthews, and Evan Freitag. 2020. Quantifying Gender Bias in Different Corpora. In *Companion Proceedings of the Web Conference 2020*, pages 752–759, New York, NY, USA. ACM.

Pablo Badilla, Felipe Bravo-Marquez, and Jorge Pérez. 2020. WEFE: The word embeddings fairness evaluation framework. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 2021-January, pages 430–436. International Joint Conferences on Artificial Intelligence.

Katherine Bailey and Sunny Chopra. 2018. Few-shot text classification with pre-trained word embeddings and a human in the loop. *arXiv preprint arXiv:1804.02063*.

Geetanjali Bihani and Julia Taylor Rayz. 2020. Model choices influence attributive word associations: A semi-supervised analysis of static word embeddings. In *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 568–573. IEEE.

Su Lin Blodgett, Solon Barocas, Hal Daumé III au2, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 1004–1015.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tolga Bolukbasi, Kai Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016a. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4356–4364. Neural information processing systems foundation.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016b. Quantifying and reducing stereotypes in word embeddings. *arXiv preprint arXiv:1606.06121*.

Angana Borah, Manash Pratim Barman, and Amit Awekar. 2021. Are Word Embedding Methods Stable and Should We Care About It?

Aylin Caliskan. 2017. Beyond Big Data: What Can We Learn from AI Models? In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 1–1, New York, NY, USA. ACM.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Mansi Chugh, Peter A. Whigham, and Grant Dick. 2018. Stability of word embeddings using word2vec. In *AI 2018: Advances in Artificial Intelligence*, pages 812–818, Cham. Springer International Publishing.

Clare Arrington. 2019. Assessing Bias Removal from Word Embeddings. *Student Research Submissions*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Marta R. Costa-jussà and Adrià de Jorge. 2020. Fine-tuning neural machine translation on gender-balanced datasets. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 26–34, Barcelona, Spain (Online). Association for Computational Linguistics.

Kate Crawford. 2017. The trouble with bias - nips 2017 keynote - kate crawford nips2017.

Sunipa Dev, Tao Li, Jeff Phillips, and Vivek Srikumar. 2019. On Measuring and Mitigating Biased Inferences of Word Embeddings. *34th AAAI Conference on Artificial Intelligence, AAAI 2020*, 34(05):7659–7666.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2020. Oscar: Orthogonal subspace correction and rectification of biases in word embeddings. *arXiv preprint arXiv:2007.00049*.

Sunipa Dev and Jeff M. Phillips. 2019. Attenuating bias in word vectors. *CoRR*, abs/1901.07656.

Amna Dridi, Mohamed Medhat Gaber, R Azad, and Jagdev Bhogal. 2018. k-nn embedding stability for word2vec hyper-parametrisation in scientific text. In *International Conference on Discovery Science*, pages 328–343. Springer.

Y Du, Y Wu, and M Lan. 2020. Exploring human gender stereotypes with word association test. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 6133–6143, Department of Computer Science and Technology, East China Normal University, China. Association for Computational Linguistics.

Yuhao Du and Kenneth Joseph. 2020. MDR Cluster-Debias: A Nonlinear WordEmbedding Debiasing Pipeline. *13th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, SBP-BRiMS 2020*, 12268 LNCS:45–54.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. 2010. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(19):625–660.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding Undesirable Word Embedding Associations. *57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, pages 1696–1705.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 1905–1925.

Scott Friedman, Sonja Schmer-Galunder, Anthony Chen, and Jeffrey Rye. 2019a. Relating word embedding gender biases to gender gaps: A cross-cultural analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 18–24.

Scott Friedman, Sonja Schmer-Galunder, Jeffrey Rye, Robert Goldman, and Anthony Chen. 2019b. Relating Linguistic Gender Bias, Gender Values, and Gender Gaps: An International Analysis.

Niklas Friedrich, Anne Lauscher, Simone Paolo Ponzetto, and Goran Glavaš. 2021. Debie: A platform for implicit and explicit debiasing of word embedding spaces. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 91–98.

Xianghua Fu, Wangwang Liu, Yingying Xu, and Laizhong Cui. 2017. Combine hownet lexicon to train phrase recursive autoencoder for sentence-level sentiment analysis. *Neurocomputing*, 241:18–27.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 115(16):E3635–E3644.

Bhavya Ghai, Md Naimul Hoque, and Klaus Mueller. 2021. WordBias: An Interactive Visual Tool for Discovering Intersectional Biases Encoded in Word Embeddings. *2021 CHI Conference on Human Factors in Computing Systems: Making Waves, Combining Strengths, CHI EA 2021*.

Yoav Goldberg and Graeme Hirst. 2017. *Neural Network Methods in Natural Language Processing*. Morgan; Claypool Publishers.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021a. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021b. Intrinsic Bias Metrics Do Not Correlate with Application Bias. *Proceedings ofthe 59th Annual Meeting ofthe Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 1926–1940.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2019*, 1:609–614.

Anthony G. Greenwald, Debbie E. McGhee, and Jordan L.K. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464–1480.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In

Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pages 122–133.

Prakhar Gupta and Martin Jaggi. 2021. Obtaining Better Static Word Embeddings Using Contextual Embedding Models. In *Proceedings ofthe 59th Annual Meeting ofthe Association for Computational Linguistics*, pages 5241–5253.

E O Gyamfi, Y Rao, M Gou, and Y Shao. 2020. Deb2viz: Debiasing gender in word embedding data using subspace visualization. In *11th International Conference on Graphics and Image Processing, ICGIP 2019*, volume 11373, School of Information and Software Engineering, University of Electronic Science and Technology of China Chengdu, Sichuan, 610054, China. SPIE.

Johannes Hellrich and Udo Hahn. 2016. Bad Company—Neighborhoods in neural embedding spaces considered harmful. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2785–2796, Osaka, Japan. The COLING 2016 Organizing Committee.

Max Hort, Maria Kechagia, Federica Sarro, and Mark Harman. 2021. A survey of performance optimization for mobile applications. *IEEE Transactions on Software Engineering*, pages 1–1.

Hwiyeol Jo and Ceyda Cinarel. 2019. Delta-training: Simple semi-supervised text classification using pretrained word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3458–3463.

D Jonauskaite, A Sutton, N Cristianini, and C Mohr. 2021. English colour terms carry gender and valence biases: A corpus study using word embeddings. *PLoS ONE*, 16(6 June).

Daniel Jurafsky and James H. Martin. 2020. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, 3rd edition draft. Https://web.stanford.edu/ jurafsky/slp3/.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving Debiasing for Pre-trained Word Embeddings. *57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, pages 1641–1650.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266.

Saket Karve, Lyle Ungar, and João Sedoc. 2019. Conceptor debiasing of word representations evaluated on weat. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 40–48.

Barbara Kitchenham. 2004. Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26.

Vaibhav Kumar and Tenzin Singhay Bhotia. 2020. Fair embedding engine: A library for analyzing and mitigating gender bias in word embeddings. *arXiv preprint arXiv:2010.13168*.

Vaibhav Kumar, Tenzin Singhay Bhotia, Vaibhav Kumar, and Tanmoy Chakraborty. 2020. Nurse is Closer to Woman than Surgeon? Mitigating Gender-Biased Proximities in Word Embeddings. *Transactions of the Association for Computational Linguistics*, 8:486–503.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172.

Anne Lauscher and Goran Glavaš. 2019. Are we consistently biased? multidimensional analysis of biases in distributional word vectors. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019)*, pages 85–91.

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2019. A General Framework for Implicit and Explicit Debiasing of Distributional Word Vector Spaces. *34th AAAI Conference on Artificial Intelligence, AAAI 2020*, 34(05):8131–8138.

Anne Lauscher, Rafik Takieddin, Simone Paolo Ponzetto, and Goran Glavaš. 2020. Araweat: Multidimensional analysis of biases in arabic word embeddings. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 192–199.

Edward Lee. 2020. Gender bias in dictionary-derived word embeddings. Technical report.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Yang Li and Tao Yang. 2018. *Word Embedding for Understanding Natural Language: A Survey*, pages 83–104. Springer International Publishing, Cham.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W. Black. 2019. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2019*, 1:615–621.

William Martin, Federica Sarro, Yue Jia, Yuanyuan Zhang, and Mark Harman. 2017. A survey of app store analysis for software engineering. *IEEE Transactions on Software Engineering*, 43(9):817–847.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders. *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2019*, 1:622–628.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*. International Conference on Learning Representations, ICLR.

Harshit Mishra. 2020. Reducing Word Embedding Bias Using Learned Latent Structure. In *AI for Social Good Workshop*.

Kevin Patel and Pushpak Bhattacharyya. 2017. Towards lower bounds on number of dimensions for word embeddings. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 31–36, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 1532–1543. Association for Computational Linguistics (ACL).

Bénédicte Pierrejean and Ludovic Tanguy. 2018. Towards qualitative word embeddings evaluation: Measuring neighbors variation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 32–39, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Radomir Popović, Florian Lemmerich, and Markus Strohmaier. 2020. Joint Multiclass Debiasing of Word Embeddings. *25th International Symposium on Methodologies for Intelligent Systems, ISMIS 2020*, 12117 LNAI:79–89.

XiPeng Qiu, TianXiang Sun, YiGe Xu, YunFan Shao, Ning Dai, and XuanJing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Yafeng Ren, Ruimin Wang, and Donghong Ji. 2016. A topic-enhanced word embedding for twitter sentiment classification. *Information Sciences*, 369:188–198.

Argentina Anna Rescigno, Eva Vanmassenhove, Johanna Monti, and Andy Way. 2020. A case study of natural gender phenomena in translation. a comparison of google translate, bing microsoft translator and deepl for english to italian, french and spanish. In *CLiC-it*.

Seyed Mahdi Rezaeinia, Rouhollah Rahmani, Ali Ghodsi, and Hadi Veisi. 2019. Sentiment analysis based on improved pre-trained word embeddings. *Expert Systems with Applications*, 117:139–147.

Thalea Schlender and Gerasimos Spanakis. 2020. 'thy algorithm shalt not bear false witness': An evaluation of multiclass debiasing methods on word embeddings. In *Benelux Conference on Artificial Intelligence*, pages 141–156. Springer.

Seungjae Shin, Kyungwoo Song, JoonHo Jang, Hyemi Kim, Weonyoung Joo, and Il-Chul Moon. 2020. Neutralizing gender bias in word embeddings with latent disentanglement and counterfactual generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3126–3140.

Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. 2019. Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 66–76, Hong Kong, China. Association for Computational Linguistics.

Maximilian Spliethöver and Henning Wachsmuth. 2020. Argument from old man's view: Assessing social bias in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 76–87.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. *CoRR*, abs/1906.02243.

Adam Sutton, Thomas Lansdall-Welfare, and Nello Cristianini. 2018. Biased embeddings from wild data: Measuring, understanding and removing. In *International Symposium on Intelligent Data Analysis*, pages 328–339. Springer.

Chris Sweeney and Maryam Najafian. 2019. A transparent framework for evaluating unintended demographic bias in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1662–1667, Florence, Italy. Association for Computational Linguistics.

Chris Sweeney and Maryam Najafian. 2020. Reducing Sentiment Polarity for Demographic Attributes in Word Embeddings using Adversarial Learning. In *3rd ACM Conference on Fairness, Accountability, and Transparency, FAT* 2020*, pages 359–368, MIT, Cambridge, MA, United States. Association for Computing Machinery, Inc.

Yi Chern Tan and L. Elisa Celis. 2019. Assessing Social and Intersectional Biases in Contextualized Word Representations. *33rd Annual Conference on Neural Information Processing Systems, NeurIPS 2019*, 32.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565, Baltimore, Maryland. Association for Computational Linguistics.

Francisco Vargas and Ryan Cotterell. 2020. Exploring the Linear Subspace Hypothesis in Gender Bias Mitigation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2902–2913.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, Rebekah Tromble, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, et al. 2021. Introducing cad: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303. Association for Computational Linguistics.

Ivan Vulic, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic? *CoRR*, abs/2004.04070.

Angelina Wang and Olga Russakovsky. 2021. Directional bias amplification. *CoRR*, abs/2102.12594.

Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong. 2020. Double-Hard Debias: Tailoring Word Embeddings for Gender Bias Mitigation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5443–5453, Stroudsburg, PA, USA. Association for Computational Linguistics.

Laura Wendlandt, Jonathan K. Kummerfeld, and Rada Mihalcea. 2018. Factors influencing the surprising instability of word embeddings. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Claes Wohlin. 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, EASE '14, New York, NY, USA. Association for Computing Machinery.

Zekun Yang and Juan Feng. 2019. A Causal Inference Method for Reducing Gender Bias in Word Embedding Relations. *34th AAAI Conference on Artificial Intelligence, AAAI 2020*, 34(05):9434–9441.

Haiyang Zhang, Alison Sneyd, and Mark Stevenson. 2020. Robustness and reliability of gender bias assessment in word embeddings: The role of base pairs. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 759–769.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning Gender-Neutral Word Embeddings. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 4847–4853.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining Gender Bias in Languages with Grammatical Gender. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, pages 5276–5284.

A Repository Search

Repository	Data Fields
ACM	Publication title, abstract, keywords
arXiv	All
Google Scholar	In the title with exact phrase
IEEE	All metadata
Science Direct	Title, abstract or author-specified keywords
Scopus	TITLE-ABS-KEY

Table 5: Data Fields Used during Repository Search

B WEAT Target and Attribute Sets

Test	Target Sets	Attribute Sets
3	European American names vs African American names (5)	Pleasant vs Unpleasant (5)
4	European American names vs African American names (7)	Pleasant vs Unpleasant (5)
5	European American names vs African American names (7)	Pleasant vs Unpleasant (9)
6	Male names vs Female names	Career vs Family
7	Math vs Arts	Male terms vs Female Terms
8	Science vs Arts	Male terms vs Female Terms
10	Young people's names vs Old people's names	Pleasant vs Unpleasant (9)

Table 6: WEAT tests used in this study. Number 5, 7 and 9 next to the set refer to the sources (Caliskan et al., 2016) used to define the word list in their paper. The names in Test 3 differ from those in Test 4.

C Paper Selection Result

Model	Reference	Year	Venue
GloVe	(Bolukbasi et al., 2016a)	2016	NIPS
	(Garg et al., 2018)	2018	PNAS
	(Sutton et al., 2018)	2018	IDA
	(Lauscher et al., 2019)	2019	AAAI
	(Yang and Feng, 2019)	2019	AAAI
	(Lauscher and Glavaš, 2019)	2019	SemEval
	(Karve et al., 2019)	2019	ACL
	(Kaneko and Bollegala, 2019)	2019	ACL
	(Clare Arrington, 2019)	2019	UMW
	(Spliethöver and Wachsmuth, 2020)	2020	ArgMining
	(Guo and Caliskan, 2021)	2020	AAAI
	(Wang et al., 2020)	2020	ACL
	(Vargas and Cotterell, 2020)	2020	EMNLP
	(Popović et al., 2020)	2020	ISMIS
	(Shin et al., 2020)	2020	EMNLP
	(Kumar and Bhotia, 2020)	2020	ACL
	(Dev et al., 2020)	2020	arXiv
	(Lee, 2020)	2020	Stanford
	(Mishra, 2020)	2020	CRCS
	(Du and Joseph, 2020)	2020	SBP-BRiMS
	(Bihani and Rayz, 2020)	2020	WI-IAT
	(Du et al., 2020)	2020	EMNLP-IJCNLP
	(Sweeney and Najafian, 2020)	2020	FAT
	(Schlender and Spanakis, 2020)	2020	BNAIC
	(Borah et al., 2021)	2021	arXiv
	(Friedrich et al., 2021)	2021	AAAI
	(Jonauskaite et al., 2021)	2021	PLoS ONE
word2vec	(Bolukbasi et al., 2016b)	2016	ICML
	(Garg et al., 2018)	2018	PNAS
	(Karve et al., 2019)	2019	ACL
	(Clare Arrington, 2019)	2019	UMW
	(Schlender and Spanakis, 2020)	2020	BNAIC
	(Sweeney and Najafian, 2020)	2020	FAT
	(Wang et al., 2020)	2020	ACL
	(Vargas and Cotterell, 2020)	2020	EMNLP
	(Popović et al., 2020)	2020	ISMIS
	(Zhang et al., 2020)	2020	AACL-IJCNLP
	(Lee, 2020)	2020	Stanford
	(Du et al., 2020)	2020	EMNLP-IJCNLP
	(Bihani and Rayz, 2020)	2020	WI-IAT
	(Gyamfi et al., 2020)	2020	ICGIP
	(Borah et al., 2021)	2021	arXiv
	(Ghai et al., 2021)	2021	CHI EA
fastText	(Lauscher et al., 2019)	2019	AAAI
	(Lauscher and Glavaš, 2019)	2019	SemEval
	(Karve et al., 2019)	2019	ACL
	(Clare Arrington, 2019)	2019	UMW
	(Popović et al., 2020)	2020	ISMIS
	(Bihani and Rayz, 2020)	2020	WI-IAT
	(Borah et al., 2021)	2021	arXiv
	(Friedrich et al., 2021)	2021	AAAI
CBOW	(Lauscher et al., 2019)	2019	AAAI
	(Lauscher et al., 2020)	2020	AAAI
	(Friedrich et al., 2021)	2021	AAAI
dict2vec	(Lauscher et al., 2019)	2019	AAAI
	(Lee, 2020)	2020	Stanford
Numberbatch	(Schlender and Spanakis, 2020)	2020	BNAIC
SGNS	(Ethayarajh et al., 2019)	2019	ACL

Table 7: Studies on Standard Static Word Embedding Models.

Bias Metric	References	Year	Venue
Word Embedding Association Test (WEAT)	(Sutton et al., 2018)	2018	IDA
	(Lauscher et al., 2019)	2019	AAI
	(Lauscher and Glavaš, 2019)	2019	SemEval
	(Tan and Celis, 2019)	2019	NeurIPS
	(Karve et al., 2019)	2019	ACL
	(Gonen and Goldberg, 2019)	2019	NAACL HLT
	(Kurita et al., 2019)	2019	ACL
	(May et al., 2019)	2019	NAACL HLT
	(Ethayarajh et al., 2019)	2019	ACL
	(Schlender and Spanakis, 2020)	2020	BNAIC
	(Guo and Caliskan, 2021)	2020	AAAI
	(Wang et al., 2020)	2020	ACL
	(Vargas and Cotterell, 2020)	2020	EMNLP
	(Lee, 2020)	2020	Stanford
	(Popović et al., 2020)	2020	ISMIS
	(Du and Joseph, 2020)	2020	SBP-BRiMS
	(Shin et al., 2020)	2020	EMNLP
	(Dev et al., 2020)	2020	arXiv
	(Zhang et al., 2020)	2020	AACL-IJCNLP
	(Borah et al., 2021)	2021	arXiv
	(Friedrich et al., 2021)	2021	AAAI
SemBias	(Kaneko and Bollegala, 2019)	2019	ACL
	(Shin et al., 2020)	2020	EMNLP
	(Kumar et al., 2020)	2020	TACL
	(Mishra, 2020)	2020	CRCS
Neighbourhood Metric	(Wang et al., 2020)	2020	ACL
	(Zhang et al., 2020)	2020	AACL-IJCNLP
Direct Bias	(Babaeianjelodar et al., 2020)	2020	WWW
	(Zhang et al., 2020)	2020	AACL-IJCNLP
Double Bind	(Tan and Celis, 2019)	2019	NeurIPS
	(May et al., 2019)	2019	NAACL HLT
Angry Black Woman (ABW) Stereotype	(Tan and Celis, 2019)	2019	NeurIPS
	(May et al., 2019)	2019	NAACL HLT
ECT	(Dev et al., 2020)	2020	AAAI
	(Friedrich et al., 2021)	2021	
Indirect Bias	(Vargas and Cotterell, 2020)	2020	EMNLP
Equity Evaluation Corpus (EEC)	(Sweeney and Najafian, 2020)	2020	FAT
MAC	(Schlender and Spanakis, 2020)	2020	BNAIC
RNSB	(Schlender and Spanakis, 2020)	2020	BNAIC
Bias-by-projection	(Yang and Feng, 2019)	2019	AAAI
Contextual Embedding Association Test (CEAT)	(Guo and Caliskan, 2021)	2020	AAAI
Sentence Embedding Association Test (SEAT)	(Kaneko and Bollegala, 2021)	2021	ACL
BAT	(Friedrich et al., 2021)	2021	AAAI
IBT	(Friedrich et al., 2021)	2021	AAAIAAAI
SQ	(Friedrich et al., 2021)	2021	AAAI
RIPA	(Zhang et al., 2020)	2020	AACL-IJCNLP
RND	(Ghai et al., 2021)	2021	CHI EA
IAT	(Du et al., 2020)	2020	EMNLP-IJCNLP
K-means Accuracy	(Du and Joseph, 2020)	2020	SBP-BRiMS
SVM Accuracy	(Du and Joseph, 2020)	2020	SBP-BRiMS
Correlation Profession	(Du and Joseph, 2020)	2020	SBP-BRiMS

Table 8: Studies on Bias Metrics for Word Embeddings.